DOCUMENT RESUME

ED 081 768 SP 006 918

AUTHOR	Waller, Michael; Soltz, Donald
TITLE	The Classical Psychometric Method of Evaluation of FTPP.
INSTITUTION	Chicago Univ., Ill. Ford Training and Placement Program.
PUB DATE	Jan 73
NOTE	17p.; Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, Louisiana, February 1973
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	Beginning Teachers; College School Cooperation; *Evaluation Techniques; Performance Criteria; Program Evaluation; *Psychometrics; *Teacher Education; Teacher Improvement; *Teacher Programs; Teacher Selection
IDENTIFIERS	Competencies; Ford Training and Placement Program; *FTPP

ABSTRACT

Data collected from one high school was evaluated to exemplify how the psychometric method was applied to the Ford Training and Placement Program (FTPP). Evaluation of the FTPP relied on both affective, paper-and-pencil measures and classroom observations. It was the concern of this investigation to discover differences between Ford and non-Ford subjects. An analysis of covariance design, with pretest scores covaried out, revealed few significant differences in posttest scores. A difference in three personality measures seemed to support the validity of the teacher selection process. The lack of other differences between Ford and non-Ford individuals may not reflect shortcomings on the program so much as the inappropriateness of the measures used. A tighter connection between program and evaluation would have made an adequate summative evaluation possible. (Author/JA)

THE CLASSICAL PSYCHOMETRIC METHOD OF EVALUATION OF FTPP

by

Michael Waller
University of Chicago

Donald Soltz
University of Chicago

January, 1973

For presentation at
the American Educational
Research Association Meeting.
February, 1973

THE CLASSICAL PSYCHOMETRIC METHOD OF EVALUATION OF FTPP

By M. Waller and D. Soltz

This paper will describe and attempt to evaluate the data collected from one high school as a means of exemplifying how the psychometric method was applied to the Ford Program. It should be noted that only the final year of operation is considered here. Hence, this evaluation of FTPP differs from those made in previous years, since earlier evaluations were seen as opportunities to experiment and re-evaluate successes and failures in reaching a number of goals. Those "formative" evaluations were intended to "improve" rather than "prove." The final year's evaluations were what the former director of research termed "summative." [1] The differences in the two types of evaluations are not so much in what is measured as what use the measures are put to. It is hoped that these quantitative data afford at least a partial summary of the FTPP.

The conceptual basis, choice of measures used, and original plan for evaluation all derive from Wayne Doyle's work and writings. In particular, his 'Transactional Evaluation in Program Development [2] contains a thorough description of his rationale for choosing the particular measures he did to evaluate the FTPP. The present paper will report some partial results from analysis of the quantitative, classical, paper-and-pencil instruments and observational data. But first, we should like to indicate that the breadth of goals articulated for the program and the possible effects of reaching those goals on the teachers, students, and others involved far outran the

ability of classical psychometric methods alone to adequately measure them. For example, six major goals and a number of related, minor ones are presented here to give some indication not only of the scope of the program, but of the formidable task of evaluation as well. One immediate limit to evaluation is thus the number of variables that might be dealt with. Some judicious parsing of possibilities for dependent variables had to be accomplished.

The first goal, developing competencies in the teacher, involved such possibilities for measurement as self-knowledge and knowledge of the learning processes; the development of skills; an increased comprehension of Black culture; knowledge of the learner as a social and academic entity; knowledge about the classroom. Other competencies included understanding the school as a social system; knowledge of the community and the environment that surrounds the school; knowledge of the process of change, and of his subject matter; knowledge of how to be an effective group member, and how to identify and define problems; and the development of problem analysis and decision-making skills.

A second goal of the program was to facilitate the induction of newly trained professionals into the school community. They are to understand their role in the cadre as well as other roles in the school and the community; know human and material resources available; accept others in the school and community and interact with them.

The third goal was to promote close relations between schools and community: knowledge of the community and its resources; awareness of the communities' expectations and problems; and the promotion of mutual participation of the community in the school.

The fourth goal of the program was the development of programs appropriate to specific classes of cadre members. This entails "new and appropriate curricula."

The fifth goal of the program involved the identification of school and community problems and to act on them within the professional competence.

Finally, there was the goal of developing an "aura of shared responsibility" among the cadre, other staff, and the community for the educational program of the school. Long-term employment, increased communication, encouragement of collegial relations, faculty decisions on matters important to it, the promotion of accountability, and the mutual support of school and community are involved.

It is clear that this impressive list of items is difficult to evaluate as a piece. Indeed, a number of different measures are necessary to make sense of the possibilities for evaluation inherent in the list. We would be interested in learning of any program that managed to achieve so many goals, but such evidence is difficult to come by. Even without hard data, however, it is clear that many of these goals have been met. Curricula have been developed specific to the FTPP, the communities have been involved, and many other goals reached. But the difficulties in measuring, say community participation by cadre members are large. There is also a risk that whatever is measured will not be accepted by one's peers as significant data. We could have counted, for example, how many non-school meetings in the community the Ford teachers attended, but how much this would tell us about the quality of educational

practice within the school is unclear. Nevertheless, as close relations between the schools and communities were clearly an expected outcome of the original model, such measures would have provided data for evaluating that aspect of the program's application of the model. A good deal of data of this sort may be garnered from the case studies and other non-quantitative materials of the program, but the question of their effect on the children and their teachers may have to be addressed more directly to be of value or even answerable.

Some difficult questions are raised for the evaluater intent on obtaining quantitative data. Should he discard the traditional achievement data on the grounds that it is irrelevant or will show no effect over the short run? This would mean sacrificing one of the most widely accepted, easily measured set of variables. Furthermore, measures in the "affective domain" (as Bloom would term those actually used) tend to have less desirable measurement properties than "cognitive" data. Reliabilities are often small due to large error variances, and validity is often a nightmarish consideration that involves thorny questions of interpretation, subjectivity of items, difficult comparisons of subject groups and so forth. Doyle determined that the most likely place for measurable quantitative changes to emerge were in terms of how the teachers and students viewed their classrooms and themselves. Hence, non-cognitive "affective" variables were used exclusively to evaluate the degree to which FTPP affected the children in classrooms of Ford-trained teachers and to view the differences in Ford versus non-Ford personnel.

A major concern for this paper is an evaluation of Doyle's primary question: "What difference does the Ford approach to training and placing professionals make in the lives of the ultimate client--the youngster in the classroom?" (p. 195)[3]. This is clearly an important question, for even if all of the goals of training and placement were met satisfactorily but no differences in students were forthcoming, the program would seem at least a partial failure. The question is, from an evaluational viewpoint, what changes might be expected, measurable, and most significant in terms of the program and the child? The nature of the effects FTPP should have had on inner-city children are implied in the 1967 article[4] containing practical proposals. In particular, some effect on the achievement of the children might not only be expected but desirable. However, Doyle has argued that the cognitive growth of children touched by FTPP may not be apparent over the short term[5]. A delayed effect extending over several years might show positive gains. Besides, cognitive gains are not clearly an explicit goal of the Ford Program in operation.

Design

Overall, three separate sets of data were obtained. 1) Measures were made of a number of Ford and non-Ford teachers' attitudes and beliefs about the nature of teaching. 2) Students were asked to respond to instruments attempting to tap their academic self-concept and their perceptions of the classrooms of which they were a part. 3) Finally, a corpus of observational data was collected from Ford and non-Ford classrooms. The overall question confronted by these data was whether or not any systematic differences existed in the way FTPP teachers and students perceived themselves and behaved as compared with a sample of non-FTPP teachers. A second question was concerned with changes related to FTPP's participation during the 1971-72 academic year. Hence, these measures were made twice, early and late in that school year.

As the Ford teachers had been selected partly on the basis of certain personality characteristics considered suitable for adequate participation in group work (as required by the nature of FTPP), three

psychological measures were also administered. These were used as independent variables in the analyses of some of the data.

In summary, then, the design called for measures of Ford and non-Ford teachers and students on a number of paper-and-pencil, "affective" measures, as well as observational data from their classrooms. These sets of data are described in detail below, but are listed here for convenient reference:

Instrumentation

I. Teacher Variables:

A. Independent Variables

Dogmatism
Flexibility/Rigidity
Psychological Distance

B. Dependent Variables

Beliefs about Research
Receptivity to Curricular Change

Teacher Conceptions of the Educational Process

Subject Matter Emphasis
Personal Adjustment
Student Autonomy
Emotional Disengagement
Consideration of Student Viewpoint
Classroom Order
Student Challenge
Integrated Learning

II. Student Variables:

A. Dependent Variables

Academic Self-Concept
Class Sentiment Index

III. Observational Variables:

	Cognitive /Memory	Critical Thinking	Expressed Emotion	Mgmt.
Seeking				
Informing				
Accepting				
Rejecting				

The total population consisted of 48 secondary and 9 elementary classrooms. The analyses have been proceding with the data from all schools, but the results given here will focus on the King school.

The design is that of a pre- post-test of two sub-groups from an essentially homogeneous population. Thus, the indicated analysis is through the use of analysis of covariance (Ancova), examining the post-test scores with the pre- test covaried out.[6] With respect to the King school data and for the purposes of this paper, the one hypothesis tested was that of treatment, in which participation in FTPP is evaluated against a comparison group of non-Ford teachers and students.

The argument is made that if the Ford Program is to be considered "effective", Ford teachers, by the end of the academic year, should differ from those teachers who have not been exposed to the same training experience in their beliefs about education. One set of questions involved a comparison of Ford and non-Ford teachers as to certain perceptions of classroom activities and their reactions to them. In particular, Ford trained teachers:

> . . .should perceive greater value in classroom research, be more receptive to curricular change, not place as much importance on subject matter, be more oriented toward the personal adjustment of students, stress the importance of the students' viewpoint, place less importance on maintaining classrooms according to established rules and procedures, and place greater emphasis on integrating content with the broader aspects of the students' world, so that the children are aware, not only of the facts of what they learn, but also of the meaning. (p. 216).[7]

In order to evaluate whether the above assertions were supportable, the administration of a battery of different measures to teachers was necessary. These are briefly described here. First, a five-item scale, termed "Beliefs About Research", was given to all teachers in the sample.[8] It requires the respondent to indicate on a five-point scale ranging from "strongly agree" to "strongly disagree" his responses to such items as "There is little research done that is of value to classroom teachers."

A second scale originated and standardized by Wehling and Charters termed "Teacher Conceptions of the Educational Process"[9] and including eight factored scales.

1) Subject Matter Emphasis, a dimension which measured the teacher's relative attribution of importance to more traditional subject matters.

2) Personal Adjustment Ideology, measuring the degree to which the teacher indicates a concern with students' problems.

3) Emotional Disengagement, a measure of social distance between teacher and student.

4) Student autonomy vs Teacher Direction, a bipolar factor dealing with the nature of the teacher's management of the classroom.

5) Consideration of the Student Viewpoint, a measure of empathy.

6) Classroom Order, measuring the degree to which the teacher perceives learning as taking place best in an atmosphere of order and decorum.

7) Student Challenge, a factor measuring the degree to which learning should be induced in students.

8) Integrative Learning, a measure of the belief by the teacher in the students' achieving emotional understanding in learning.

The third instrument used was one drafted by Bridges, called "Teacher Receptivity to Change."[10] He concluded that his seven items validly discriminated between teachers willing to undertake an innovative practice as against those not so inclined.

With respect to differences in students taught by Ford as opposed to non-Ford teachers, Doyle [11] argued that:

> From the students' point of view, the teaching practices should be different (i.e., in Ford-taught rooms versus non-Ford). The students in Ford classrooms should judge the work to be more exciting. Ford teachers should use praise more often to reinforce students positively; they should encourage students more often to express feelings about the teaching. The teachers should encourage students to express ideas that are different from their own, they should

> talk less than other teachers, should encourage students to lead discussions, should assign individual and group projects more often, use small groups for special projects, and, in examinations, should not test facts and memories exclusively. The students in Ford classrooms should consider that their teachers have less rigid expectations for them and are more "creative and open". (216-217)

Students were asked to respond to two paper-and-pencil instruments measuring three variables. Each test was composed of items developed at the Instrumental Objectives Exchange in Los Angeles. Items for the first instrument, the Class Sentiment Index (CSI)[13] were selected to measure two variables; the students' perceptions of their teachers' teaching behavior; and the students' perceptions of the teachers' feelings about themselves. We called the second instrument the Academic Self-Concept Inventory (ASCI).[12] Items for the ASCI were selected so as to measure the students' self-concept in the context of his class.

For each instrument, extensive item analysis was performed utilizing Lazarsfeld's Latent Structure or Latent Trait Analysis.[14] In this particular application of Lazarsfeld's theory, the form of the function relating the underlying trait to the probability of a positive response on any item was assumed to be the Normal Ogive. In Latent Trait Analysis the parameters of this function are estimated and a test of fit produced for each item as well as for the instrument as a whole. When an item doesn't fit the model it is assumed that the responses to that item were motivated by more than the one variable intended. In this way items which do not appear to measure the same trait as the set of items as a whole can be eliminated and one is therefore able to obtain an instrument which is measuring essentially one underlying trait. The results of this analysis are presented below.

A third set of data was obtained from non-participant observers using an objective observational instrument called CERLI Verbal-Behavior Classification System (CVC).[15] Doyle predicted that

the observers "should have noted differing classroom practices" in Ford teachers' rooms as compared with non-Ford rooms. Explicitly, non-participant observers should have noted "differing classroom practices," as follows: "Ford teachers talk less than other teachers, use praise to reinforce students positively, seek fewer responses that call for critical thinking " (217) [16] Finally, students in Ford-taught classes might well give less factual information, more critical thought, express more positive sentiment toward their classes and higher estimates of their own academic performance.

The CVC allows an observer to tally the percentage of time teacher and student talk. In addition, all "talk" can be further broken down to fit into one or a combination of 16 cells so that its qualities can be examined more precisely. In particular, all verbal activity in the classroom can be coded according to four content categories--Cognition memory; productive, critical thinking; expressed emotion; management -- and four categories of process used -- seeking, informing, accepting, rejecting. That is, each verbal exchange can be classified most grossly as to whether it was emitted by teacher or pupil, then more finely as to which of 16 possible combinations of process and content it belongs to. Thus it becomes possible to test for differences in Ford and non-Ford classes as to which ones contain more teacher talk, seeking behavior, praising of students, more critical thought, and so forth.

For the purposes of this report, only six classrooms at Martin L. King High School were available for analysis. Two Ford teachers and four non-Ford teachers were observed on several occasions both early and late in the school year. Comparisons on all categorizations of verbal behavior of interest to this study revealed no significant differences either in time when measures were made or in Ford vs non-Ford comparisons.

Independent Variables

Effective participation in the Ford Program requires, among other things, a personality suitable for close work in groups, and one which is open to innovative change in one's teaching. Accordingly, the selection process used by the Ford staff hopefully would be expected to produce a group of teachers with personality characteristics different from the general population of teachers. In order to test for

the effectiveness of the selection process (and to get a handle on possible sources of bias in the study) all Ford teachers and a selection of non-Ford teachers at each school were measured on three personality traits. It was assumed that their ability to work in groups might be indicated by their scores on the Psychological Distance Scale, [17] and their openness to change by their scores on the Rokeach Dogmatism Scale E [18] and the Flexibility/Rigidity Scale.[19]

Table I

Test of Group Differences

(Ford vs. Non-Ford) on the Paper-and-Pencil Measures

	Univariate P	Multivariate P
Paper and Pencil Measures		
Independent variables		
Dogmatism	NS	.05
Flexibility/Rigidity	NS	
Psychological Distance	.01	
Teachers		
Dependent variable		
Receptivity to Curriculum Change	NS	NS
Attitude Towards Research	.05	
Subject Matter Emphasis	NS	
Personal Adjustment	NS	
Student Autonomy	NS	
Emotional Disengagement	NS	
Consideration of Student Viewpoint	NS	
Classroom Order	NS	
Student Challenge	NS	
Integrated Learning	NS	
Students		
Dependent Variable		
CSI 1. Teaching Behavior	NS	NS
2. Feelings About Students	NS	

Results and Discussion

As we consider Table 1, we note first that the teachers selected for the Ford Program did in fact differ significantly on the three personality measures taken together from those non-Ford teachers on which we obtained data. However, before considering this result further,

we wish to point out that there is a problem in its generalization. Biased selection of participants which precludes the basic assumption of randomization underlying the general linear hypothesis (e.g., ANCOVA) severely limits the generalizability of results. Simply selecting teachers in a manner such as that used in the Ford Training and Placement Program--one designed to produce a group of teachers open to innovative change--and then giving this group a certain sum of money to produce change might easily explain any differences in the students. Of course, we could (and, in fact, did, with no change in the results) covary out these independent variables in an attempt to control for this selection bias. But we are still in a situation of limited generalizability. Furthermore, other factors relevant to the measured outcomes, but unknown to the researcher, may be present, particularly in the affective domain.

The significant difference reported in Table I between Ford and non-Ford teachers on personality measures would seem to support the contention that the selection process used in the Ford Training and Placement Program produced a group of teachers with characteristics different from the general population of teachers. However, the non-Ford teachers were self-selected in that they agreed to be so measured. This is again a biased sample from the relevant population (this time of non-Ford teachers) and in terms of confirmatory data analysis is again equivocal. However, we would argue that removing this source of bias (i.e., including teachers who were approached, but refused to participate) would only effect the result obtained on the Dogmatism Scale, the Flexibility/Rigidity Scale, or the Psychological Distance by increasing the observed difference. Therefore, we conclude that the data collected tend to support the hypothesis that the selection process produces teachers with characteristics different from the general population of teachers. The result on the Psychological Distance Scale would support particularly the contention that the selection process produces a group of teachers who possess at least one characteristic deemed desirable for group work.

Of course, one can disagree with this argument (and we welcome such comments); however, this possibility of disagreement

allows us to state our primary point with respect to this issue. The only way out of such an equivocal situation is to randomize. In this context we simply note that there was one dependent variable which showed the expected result: Ford teachers' responses indicated a more positive attitude toward research than non-Ford teachers' responses. As the training sessions and many other activities of the cadres were closely related to the strongly research oriented University of Chicago, such a result is not surprising.

With respect to the students, we are in a much better position, at least in terms of randomization. While we might not be able to attribute any worthwhile results to the Ford Program as such, the treatment was applied to a random selection of students, and the treatment, selection process, or whatever, is essentially replicable. However, the question of generalizability becomes moot. The student responses to the subscales of the CSI concerning the students' perceptions of two aspects of their teachers' behavior show no significant differences between students in the Ford and non-Ford classes. The Student Self-Concept Index, while yielding a very good fit in the elementary school population, proved to be inappropriate for the high school students.

In evaluating the data as collected (as opposed to the "quantitative method"), an additional comment seems in order. Paper-and-pencil tests of affective traits are notorious for large error variances. For example, in all three student variables the range of scores was slightly less than twice the standard error. Therefore, while we clearly are not in a position to reject the null hypothesis of no difference between groups, the imprecise nature of these measures motivates the statisticians' primary admonition: one does not accept the null hypothesis, one fails to reject the null hypothesis.

Discussion

It should be clear by now that the few differences in the Ford teachers and students as compared with non-Ford may mean many things: inappropriate measures, unreliability of those used, too small a sample and so forth. One thing seems clear, however. The analyses of the FTPP as reported in this paper are wrong to the degree that they indicate no changes taking place. The program wrought

many and profound changes, as the other modes of evaluation indicated. Of this there can be no doubt. Perhaps our shortcoming is that we, as psychometricians, have been wedded to the traditional paper-and-pencil approach to evaluation. It is simple enough to distribute a group of already constructed tests and hope for some differences to show up. Then we can claim that the program was the cause. It is far more difficult to discover those elements in the program that should result in change in the participants and to devise ways to quantitatively measure these. And further, to adhere to traditional canons of measurement and test theory in doing it. Life was simply too short to do that in the case of FTPP. As our assistant director, Jim McCampbell, is fond of saying, "The trouble with this program is that it's always a year behind itself." Indeed, when we should have been devising appropriate measures for the final evaluation, we were in the midst of one of the countless crises that determined the course of events. The purposes were--and are--still evolving. A quantitative approach that finds a rational basis in the nature of the FTPP itself is difficult in such a fluid environment. And without a firm grip on the limitations and aims of the program and, in particular, the processes by which these are to be achieved, quantitative evaluation of the sort outlined here seems inappropriate. What is required is seemingly a shifting, responsive means of securing numerical data that, at the same time, is stable enough to be useful in decision-making. We have not yet found this.

It seems clear that the evaluation needed, but lacking, in the FTPP required the formation of a closer connection between that which is measured and the program's intent. This should be implied by this summary. We can suggest some examples of the sorts of comparisons that were not made using quantitative data, but were clearly suggested by the Getzels' article.[20] The frequency of professional interactions among Ford teachers could easily have been greater than that among non-Ford teachers; paper-and-pencil tests of the teachers' knowledge of the different roles that exist in the social system of schools would have probably shown significant differences; and related to this knowledge, more effective use of the role specialists would

probably have been made by Ford teachers. To the extent that such relationships were not examined, we suppose that the FTPP is improperly represented by the present data. Only when program and evaluation are specifically and firmly linked together will an adequate summative evaluation be possible.[21]

1. Terms taken from M. Scriven, "The Methodology of Evaluation." in Perspectives of Curriculum Evaluation (Chicago: Rand MCNally, 1967).

2. W. J. Doyle, "Transactional Evaluation in Program Development," in Robert Rippey (ed.) Studies in Transactional Evaluation, in press.

3. Ibid.

4. J. W. Getzels, "Education for the Inner City: A Practical Proposal by an Impractical Theorist," School Review, 75, 1967, 283-299.

5. Doyle, op. cit.

6. R. D. Bock, "Unconditional Inference in the Analysis of Repeated Measurements" paper read at Symposium on the Application of Statistical Techniques to Psychological Research, Canadian Psychological Association, York University, June 4, 1969.

7. Doyle, op. cit.

8. This test is original to Wayne J. Doyle; copies available on request.

9. B. Wehling and W. W. Charters, "Dimensions of Teacher Beliefs about the Teaching Process," American Educational Research Journal, 6, 1969, 7-30.

10. F. Bridges, "The Measurement Problem in Teacher Receptivity to Change," unpublished mimeo, 1968.

11. Doyle, op. cit.

12. Derived from Measures of Self Concept (Los Angeles: Instructional Objectives Exchange, 1972).

13. Attitude Toward School Los Angeles: Instructional Objectives Exchange, 1972).

14. P. Lazarsfeld, "Latent Structure Analysis," in M. Fishbein (ed.) Readings in Attitude Theory and Measurement (NY: Wiley, 1967).

15. E. House, et al., The Gifted Classroom, ERIC 054 594.

16. Doyle, op. cit.

17. F. E. Fiedler, "The Leader's Psychological Distance and Group Effectiveness," in D. Cartwright and A. Zander (eds.), Group Dynamics: Research and Theory (New York: Harper and Row, 1960, 2nd ed., pp. 586-606).

18. M. Rokeach, et al. The Nature of Analysis and Synthesis, Cooperative Research Project 879 (Washington, D.C.: U. S. Printing Office, 1964).

19. Ibid.

20. Getzels, op. cit.

21. Susan S. Stodolsky, "Defining Treatment and Outcome in Early Childhood Education," in H. Walberg (ed.) Rethinking Urban Education, Jossey-Bass, in press.